

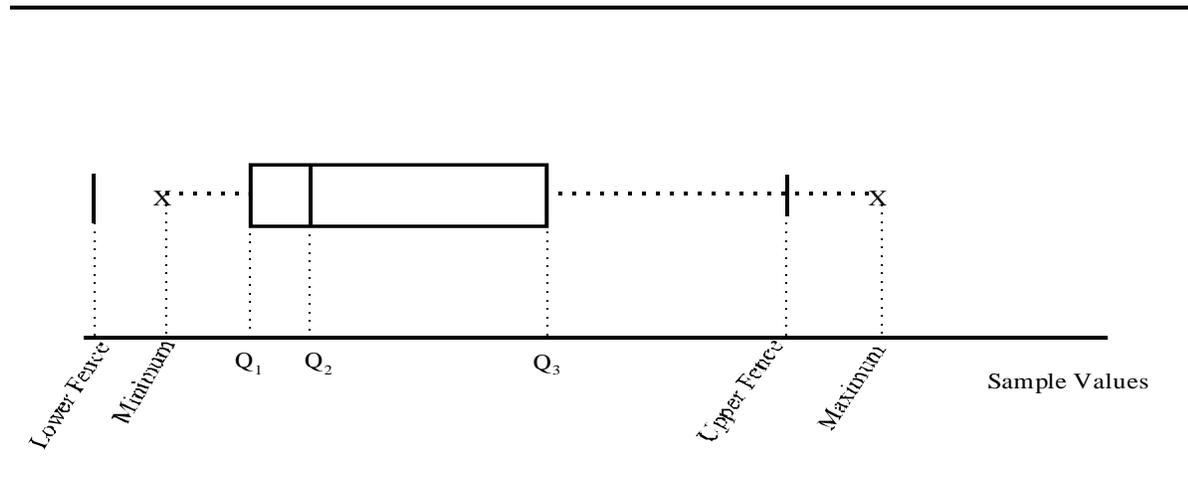
BOX-PLOT PRIMER
DOCKET NO. 02-057-02
DPU EXHIBIT 6.12

John Tukey invented the boxplot as a means of visually summarizing the characteristics of an actual sample as opposed to assuming a data generating process or distributional form. The boxplot, at a glance, shows the central location, spread, skewness, tail length, and outlying data points for the sample under investigation.

Construction of the boxplot is relatively easy, needing only a few simple sample statistics: the sample fourths or sample quartiles, and the minimum and maximum sample values. The quartiles are those values that divide the data into fourths. Specifically, the first quartile, Q_1 , is that value such that, 25% of the sample values are less than it. The second quartile, Q_2 , is defined as the sample median and is that value such that, 50% of the sample values are less than it. And the third quartile, Q_3 , is that value such that, 75% of the sample values are less than it.

The boxplot is constructed by drawing a rectangular box whose ends are the first and third quartile. The second quartile, which is the median, is depicted by a vertical line within the box at the corresponding value. The maximum and minimum values are depicted by an "x" over the respective value with a connecting line drawn from the x to the closest end of the box. Additionally, an upper and lower fence or boundary, are depicted with vertical lines over the appropriate values, are drawn in the resulting graph. The boundaries are designed to indicate the presence of outliers in the sample data: sample values greater than the upper boundary or less than the lower boundary would be considered unusual data values – data values that have a low probability of occurrence.

Figure 1: Example of a Box Plot



The upper and lower boundaries are constructed using the $1.5 \cdot \text{IQR}$ rule proposed by Tukey. The lower boundary is calculated as $Q_1 - 1.5 \cdot \text{IQR}$, and the upper boundary is calculated as $Q_3 + 1.5 \cdot \text{IQR}$, where IQR (the inner quartile range) is defined as $Q_3 - Q_1$. Supposedly, when a student asked why 1.5, Tukey replied, “Because 1 is too small and 2 is too large.” Nevertheless, it can be shown for a wide variety of distributions that, the $1.5 \cdot \text{IQR}$ rule is reasonable for determining if a sample value is an outlier.

For instance, consider the standard Gaussian distribution with a mean of zero and variance of one. For symmetric distributions the median and the mean are the same, in this case 0. And, for the standard normal distribution, the first and third quartiles are given by -0.6745 and 0.6745 respectively. Thus, the upper and lower boundaries, 2.698 and -2.698, contain about 99.3% of the distribution. In other words, if a random sample is drawn from a standard normal distribution, we would expect less than 1% the values in that sample to lie outside of the boundaries as defined by the $1.5 \cdot \text{IQR}$ rule.

Similarly, for a Chi-square distribution with five degrees of freedom, we would expect less than 3% of the sample values to lie outside of the boundaries. If there were twenty degrees of freedom, we would expect less than 1.4% of the sample values to be outside of the boundaries.¹

¹ For a more complete discussion of the use of boxplots, see John D. Emerson and Judith Strenio, “Boxplots and Batch Comparison,” in *Understanding Robust and Exploratory Data Analysis*, David C. Hoaglin, Fredrick Mosteller, and John W. Tukey, editors, (New York, John Wiley & Sons), 1983, pp.58-65. Tukey’s original work can be found in, *Exploratory Data Analysis*, Addison-Wesley, 1977.

Table 1: Comparison of the 1.5*IQR Rule

Distribution	Probability Outside of Boundaries				
	Lower Boundary	Upper Boundary	Below Lower Boundary	Above Upper Boundary	Total
Standard Normal	-2.70	2.70	0.004	0.004	0.007
Chi-Square 5 df	-3.25	12.56	0.000	0.028	0.028
Chi-Square 20 df	2.89	36.39	0.000	0.014	0.014

If a researcher wishes to use a more conservative rule, a 3*IQR rule could be used to identify outliers. Using this rule, we would expect less than 1/100 of 1% of data values to lie outside the boundaries for a standard normal distribution. Similar results are expected for the two Chi-square distributions presented here.

Table 2: Comparison of the 3*IQR Rule

Distribution	Probability Outside of Boudaries				
	Lower Boundary	Upper Boundary	Below Lower Boundary	Above Upper Boundary	Total
Standard Normal	-4.72	4.72	1.17E-06	1.17E-06	2.34E-06
Chi-Square 5 df	-9.18	18.48	0.00	2.40E-03	2.40E-03
Chi-Square 20 df	-9.68	48.96	0.00	3.12E-04	3.12E-04

While the 1.5*IQR (or the 3*IQR) rule is some what arbitrary, “experience with many data sets indicates that this definition serves well in identifying [outliers].”² And, although an analogous procedure based on the sample mean and sample standard deviation, say a 95 or 99% confidence interval, could be used, it

² John D. Emerson and Judith Strenio, “Boxplots and Batch Comparison,” in *Understanding Robust and Exploratory Data Analysis*, David C. Hoaglin, Fredrick Mosteller, and John W. Tukey, editors, (New York, John Wiley & Sons), 1983, p. 62.

would lack the necessary resistance to outliers. That is, the sample mean and standard deviation are sensitive to the presence of even one “wild” value. The quartiles, on the other hand, are resistant to the presence of a few wild data values. Indeed, “up to 25% of the data values can be made arbitrarily large . . . without greatly disturbing the median, the fourths, or the rectangular box in the boxplot.”³ Thus, we conclude, that the boxplot affords a reasonable means of identifying data values that are unusual, values that have a low probability of occurrence.

³ *Ibid.*, p. 61.